Assignment 9 - Random projections (Extra credit)

CS 675 Fall 2019

November 14, 2019

In this optional assignment we will experiment with random hyperplanes for classification. Your program will take a dataset as input and produce new features following the procedure below. The input is in the same format as for previous assignments.

Inputs:
Data matrix X: n rows, m columns
Training labels: Y
Level : k

Let Z and Z_1 be an empty list initially.
For i = 0 to k do:

(a) Create random vector w where each w_j is uniformly sampled between -1 and 1.

(b) Let x_j be our training data rows from X. Determine the largest and smallest $w^T x_j$ across all training rowsx_j in X. Select w_0 randomly in the range $[\min_j w^T x, \max_j w^T x_j]$.

(c) Project training data X (each row is datapoint x_j) onto w. Let projection vector z_i be $Xw + w_0$ (here X is $n \times m$ matrix and w is $m \times 1$ vector). Append $(1 + sign(z_i))/2$ as new column to the right end of Z. Remember that z_i is a $n \times 1$ vector and so for each row z_{ki} of z_i, $(1 + sign(z_{ki}))/2$ is 0 if $z_{ki} < 0$ and 1 otherwise.

(d) Project test data X_1 (each row is datapoint x_j^1) onto w. Let projection vector z_i^1 be $X_1 w$. Append z_i^1 as new column to the right end of Z_1.

1. Run linear SVM on Z and predict on Z_1.
2. Do for values of k=10, 100, 1000, and 10000.
3. How does the error compare to liblinear on original data X and X_1 for each k?

Submit a document containing the error of linear SVM (cross-validated C) on the first split of each of the six datasets on the course website. Do this on the original data representation and the new representation for all values of k.

Submit your program that creates features and run LinearSVC (in Python scikit) on the new training data and predicts on the new test data. In LinearSVC set the max _iter parameter to 10000 so that we do a deep search.

If you are working in a group then only one student needs to submit but both names need to be on the submission document so that we know who the team is. The other student in the team can leave their submission directory empty.

Submit your assignment by copying it into the directory
/afs/cad/courses/ccs/f19/cs/675/101/¡ucid¿.
For example if your ucid is abc12 then copy your solution into
/afs/cad/courses/ccs/f19/cs/675/101/abc12.